Incremental Training with All-Pass Transforms

John McDonough and William Byrne

Center for Language and Speech Processing

The Johns Hopkins University

May 16, 2000

Speaker Compensation

- The performance of current automatic speech recognition (ASR) algorithms degrades significantly in the presence of inter- speaker differences.
- Speaker compensation attempts to acount for or eliminate these differences and thereby improve ASR performance.
- *Speaker normalization* transforms the original cepstral features to match the speaker-independent model:

$$\hat{x}_i = T(x_i) \ \text{(normalization)}$$

- *Speaker adaptation* transforms the original cepstral means to match the features of a given speaker:

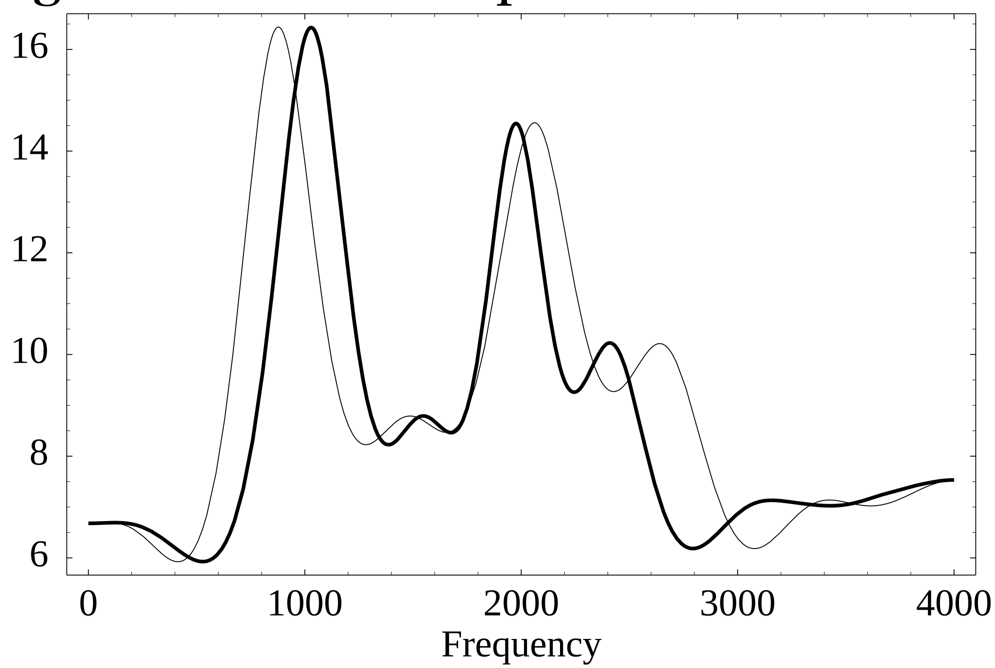
$$\hat{\mu}_k = A^{(s)}\mu_k + b^{(s)} \ \text{(adaptation)}$$

The All-Pass Transform

- The all-pass transform (APT) is a linear transformation of cepstral coefficients specified by very few free parameters (e.g., one or nine).
- In normalization, the APT warps the frequency axis associated with the short-time Fourier transform of a segment of speech (ICSLP ’98).
- In adaptation, the APT transforms the cepstral means of an HMM (ICASSP ’99).
- APT adaptation can be efficiently incorporated into HMM parameter estimation to achieve matched conditions on training and test (EuroSpeech ’99).

APT Spectral Transformation

- Original (thin line) and transformed (thick line) short-term spectra regenerated from cepstra 0–14.



- Note that the higher formants are shifted *down*, while the lowest formant is shifted *up*.

The Sine-Log All-Pass Transform

- Define the *Sine-Log All-Pass Transform* (SLAPT) as

$$Q(z) = z \exp F(z)$$

where

$$F(z) = \sum_{m=1}^{M} \alpha_m F_m(z) \text{ for } \alpha_1, \ldots, \alpha_M \in \mathbb{R},$$

$$F_m(z) = j\pi \sin\left(\frac{m}{j} \log z\right)$$

- The SLAPT shares all characteristics of RAPT, save for its rational form.
- The SLAPT, however, is more amenable for computation.

SLAPT Characteristics

- Upon applying

$$\sin z = \frac{1}{2j}\left(e^{jz} - e^{-jz}\right)$$

it follows

$$F_k(z) = \frac{\pi}{2}\left(z^k - z^{-k}\right)$$

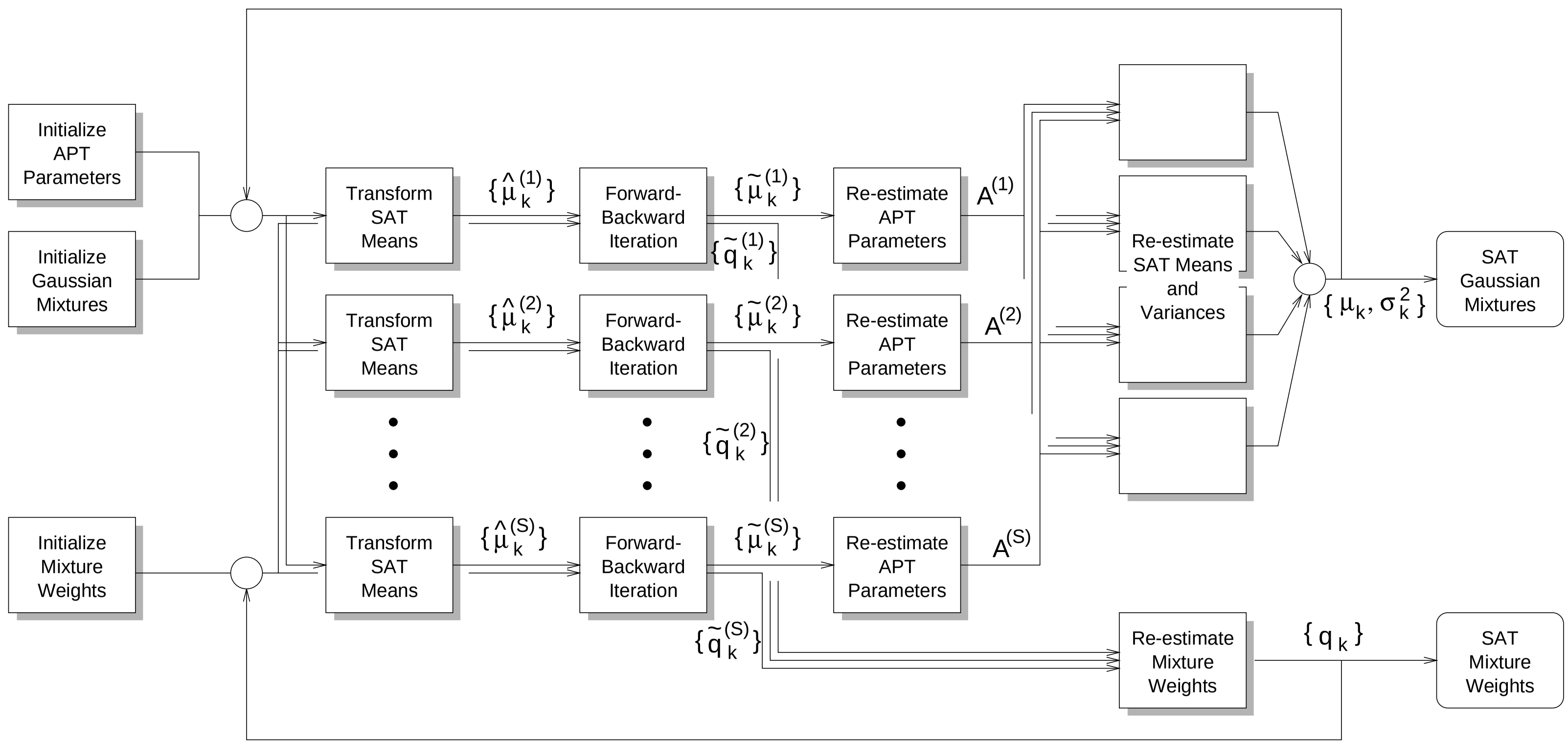
which is a better form for computation.

- Parameterizing the unit circle as $z = e^{j\omega}$ provides

$$Q(e^{j\omega}) = \exp j\left(\omega + \pi \sum_{k=1}^{K} \alpha_k \sin \omega k\right)$$

Center for Language and
Speech Processing
The Johns Hopkins University

SAT Schematic



Center for Language and
Speech Processing
The Johns Hopkins University

Multiple/Optimal Regression Classes

- In speaker adaptation, the Gaussian components of an HMM are often partitioned into mutually-exclusive sets or *regression classes.*
- An unique speaker-dependent transformation is then estimated for each regression class.
- In earlier work, the regression classes were typically obtained with binary divisive clustering or based on phonetic similarity.

Homewood Incremental Training (HIT)

The unique characteristics of the APT mandate the use of special HMM training techniques (submitted, ICASSP ’00).

1. Incrementally add speaker-dependent modeling detail to single mixture model.
2. Detail may be added by increasing the number of regression classes, or by the number of parameters per transform, or both.
3. We have developed useful heuristics for regression class splitting.
4. Modeling detail is transferred to multiple-mixture model in a computationally efficient manner.

Center for Language and
Speech Processing
The Johns Hopkins University

The Mississippi State Training Set

- Speech recognizer was trained on a subset of Switchboard Corpus training set, dubbed *MsTrain*
  - Approximately 800 conversations total;
  - Approximately 50 hours of speech;
  - Approximately 400 speakers of each gender.
- MsTrain set used in estimating a “plain vanilla” speaker-independent model:

Speaker Normalization Results

- Feature normalization was tested in combination with MLLR.
- APT parameters were estimated with a simple GMM.

Feature Normalization	Full-Matrix MLLR	
	No	Yes
None	40.6	36.3
RAPT-1	38.8	34.8
RAPT-5	39.4	35.0
SLAPT-1	38.8	34.7
SLAPT-5	39.6	35.3

- In all experiments, training and test conditions were *matched*.

APT Rapid Adaptation Results

- Sparsity of parameters in APT make it ideal for use with limited enrollment data.
- Unsupervised parameter estimation was performed using various amounts of enrollment data.

Enrollment Set	RAPT-1	RAPT-9	SLAPT-1	SLAPT-9	Full-Matrix MLLR
Baseline	41.5				
2.5 min.	38.5	37.3	38.4	37.4	37.1
60 sec.	38.3	37.4	38.2	37.5	37.5
30 sec.	38.5	37.6	38.3	37.7	37.9
10 sec.	38.7	37.8	38.6	38.0	40.1
5 sec.	38.8	37.9	38.6	38.2	45.5

- All cases used a single, global transform.

APT Adaptation Results

- The results of several speech recognition experiments using *unsupervised* APT adaptation are tabulated below.

No. Regression Classes	% Word Error Rate	
	RAPT-1	RAPT-9
Baseline	40.6	
1	38.2	37.3
2		37.0
4		36.3
8		36.1
16		36.1
24		35.6

- The use of more regression classes and more parameters per transform results in ever increasing WER reductions.
- The best WER reduction is 5.0% absolute.

MLLR Results

- Increasing the number of regression classes under MLLR yields no additional improvement.

No. RegClasses	% Word Error Rate
Baseline	40.6
1	36.9
2	36.3
4	37.3

- The best WER reduction with MLLR is 4.3%, significantly less than that obtained with APT-based adaptation.

Conclusions

- An APT-based speaker adaptation system yields an 5.0% reduction in WER on a large vocabulary conversational speech recognition task.
- The comparable gain with MLLR is 4.3%.
- Unlike conventional MLLR, the parameters of the APT can be robustly estimated in the face of limited enrollment data.
- *The Homewood Extensions* are now available at `ftp://ftp.clsp.jhu.edu/pub/the`.